KNOWN PHYLOGENIES OF ORGANISMS BASED ON THE COMPARISON OF HOMOLOGOUS E2F TRANSCRIPTION FACTORS AND HOW THESE DIFFERENT FACTORS CHANGED THROUGH EVOLUTION

A SENIOR HONORS THESIS

Presented in Partial Fulfillment of the Requirements for Graduation

With Research Distinction in Molecular Genetics in the College of Biological Science of

The Ohio State University

By

Arie H. Tan

****

The Ohio State University

Spring 2007

Honors Committee:

Dr. Gustavo Leone (Project Adviser)

Dr. Amanda Simcox (Departmental Representative)

Dr. Kun Huang (Honors Representative)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Dedicated to my parents,

whose constant love and endearing encouragement

kept me focused on my studies

Acknowledgements:

I wish to thank Professor Gustavo Leone for his guidance and patience in advising me for completing my thesis. I also wish to thank Professor Amanda Simcox and Professor Kun Huang for their contributions and their willingness to serve on my honors thesis committee. I am grateful to Dr. Pamela Wenzel for her time and effort in helping me for the completion of the analytical procedures implemented in this thesis. My special thanks go to my father for being ever so patient and supportive with my studies; to my mother, who provided me with great care and love throughout the years of my studies; and to my brother, Adrian, whose nagging for playing computer games I ignored due to my busy schedule completing my thesis (Thank god, Adrian is now a freshman at The OSU).

Vitae:

Current Grade Point Average: 3.825/4.000 (March 2004 – June 2007)

Awards and Scholarships:

- Dean's List of Honors Students (Autumn 2004 - present)
- University Scholarship (Summer 2004 - present)
- Bernstein Memorial Fund (Spring 2005)
- The 2007 Colleges of the Arts and Sciences Certificate for Excellence in Outstanding Scholarship (April 24, 2007)
- Membership in the Phi Kappa Phi Honors Society (May 2007)
- Membership in the Phi Beta Kappa Honors Society (May 22, 2007)

Volunteering and Community Service Experiences:

- Volunteer at the Dorrian Hilltop Senior Center (Summer 2006 - Autumn 2006)
- Volunteering at The Ohio State University Medical Center (Spring 2007 - present)

Abstract

This study introduces the methods used to determine the phylogeny of different E2f factors as they might have evolved over time in different organisms, diverging into the wide array shown today. It also represents a method to find new, putative E2f factors in the known genomic database through in silico analysis. To do this, techniques such as BLAST searches, sequence homology alignment, and phylogenetic comparisons were used. The findings of this study suggest that these factors may have evolved through the basis of function, although convergent evolution may be a distinct possibility in some cases. From these finds, the activator groups may have evolved separately from the repressor groups, but with different groupings than those found when the factors were independently analyzed. To make this conclusion, we also used potential factors that were found in some metazoans like Arabidopsis or Danio rerio. Further investigation may be needed to determine whether these factors may really function in the way that E2fs are known to do.

Keywords: E2f, phylogeny, BLAST search, ClustalW, activator, repressor

CHAPTER 1

INTRODUCTION

1.1 Introduction

The E2f factors are active regulators of transcription within the cell that, through the recruitment of certain cofactors, regulate whether the cell divides or not. They can be classified into three different groups: in humans, the first three (E2f1-3a) are thought to facilitate transcription. However, the other groups (which comprise of E2f3b-6 and E2f7-8) have been found to be molecular repressors of this process. The last group (E2f7-8) was discovered only recently, and was found to have a different structure of binding than the other E2f factors [Logan et al. 2005]. Also, they have highly similar layouts of domain and sometimes can even heterodimerize.

These factors have been shown to maintain a high level of conservation among particular organisms. The question remains, though: how have the constant changes wrought about by the process of evolution changed the factors from the start? More importantly, could we use these changes to make an evolutionary correlation among different organisms, and if so, how consistent is this with trees developed through morphological and/or traditionally accepted molecular phylogenetic techniques? Finally, what portions of the different sequences are conserved enough so that we could use them as a benchmark for the search of other E2f factors in organisms? This thesis presents an effort to answer these questions through analysis of the known sequences gathered throughout the scientific community.

1.2 Objective of the Study

At this point, the known E2f factors are not fully phylogenetically organized and this study is an attempt to introduce the methods to organize them. Thus, the study's main objective is to obtain the phylogeny of various known E2f factors based on their known sequential divergences.

Specifically, the first task in this research is to gather information from the scientific community for the necessary sequences of these proteins, so that a phylogeny can possibly be surmised. To compare sequences in this manner, the different E2f sequences among organisms are to be arranged by homology and once that is done, an adequate comparison can be made through phylogenetic analysis.

The next task in this research procedure is to determine how the individual factors may have diverged from a hypothetical set of primitive E2f factors that may have existed in earlier metazoans. Phylogenetic analysis will be conducted using sequences from various organisms that have been discovered so far, and the resulting data, which will be visualized through phylogram analysis via the ClustalW program, will be interpreted to achieve a hypothesis as to the evolution of the different E2f factors.

Finally, a second approach will be used in which the search for the defining character of E2f transcription factors is done, with this factor subsequently used to search for other factors through the NCBI BLAST engine.

1.3 Organization of the Report

This report is organized into three sections. First, background information as to the general function of the E2f domains is introduced, including potential mechanisms and interactions with fellow transcription factors and the cyclins. Second, the results of particular analyses of phylogenies through E2f sequences are presented, as well as the methods used to generate them in this study. Finally, the resulting implications of this information, as well as the comparisons between this and other phylogenetic trees from different analyses, are discussed.

CHAPTER 2

STATEMENT OF THE PROBLEM

2.1 Introduction

This chapter is divided into three parts. The first is the research problem statement, which explains the potential reasons for the presence of these factors and the implications of these. The second, scope and limitation, describes the limits of the particular study. Finally, the background of these proteins is shown, as well as possibilities as to how they may act in facilitating transcription.

2.2 Research Problem Statement

Our inquiry is threefold: the first aspect rests upon the correct homology of the sequences to be used in this study. Second, we need to analyze the question of the possible phylogenetic distances and relationships among different E2f factors based on analyses from the ClustalW servers. Finally, an important part of this study will involve discussion about the given results, including the resolution of any discrepancies and possible application of the data into the modern scientific knowledge. It is also important to note that we will soon need to compare it with the modern phylogenic patterns yielded from other resources and methods such as rRNA analysis, the consensus basis of genetic comparison.

2.3 Scope and Limitation

The scope of the study pertains to the protein sequences of transcription factors E2f1-8 in various organisms for phylogenetic comparisons so as to generate a hypothesis as to their relationships. The primary limitation lies within the number of model organisms sequenced and the available sequences within the known literature.

2.4 Background of Study

The characterization of the first known E2f factors was known through studies of adenoviral protein receptors; it was shown to interact with the viral E1A protein. It is now well-known that the E2f factors are very important in maintaining a check on cellular division, within the G-S checkpoint. In order for a cell to continue into synthesis, the S-cyclin must be activated, and this is accomplished through the phosphorylation of the corresponding Cdk protein. The primary mediating factor for this is the Rb kinase, which in turn must be regulated by the E2f factor in order to properly function.

As shown in Figure 2.1, the various mammalian factors are divided into three groups, numbered based on their order of discovery and characterization. The members of the first family, composed of E2f1-3, are known activators in mammalian systems, and E2f4-6 serve as repressor E2fs in said cells. The other two, having a different structure than the others, may serve as repressors in the cell, although further analysis may be needed for full characterization.

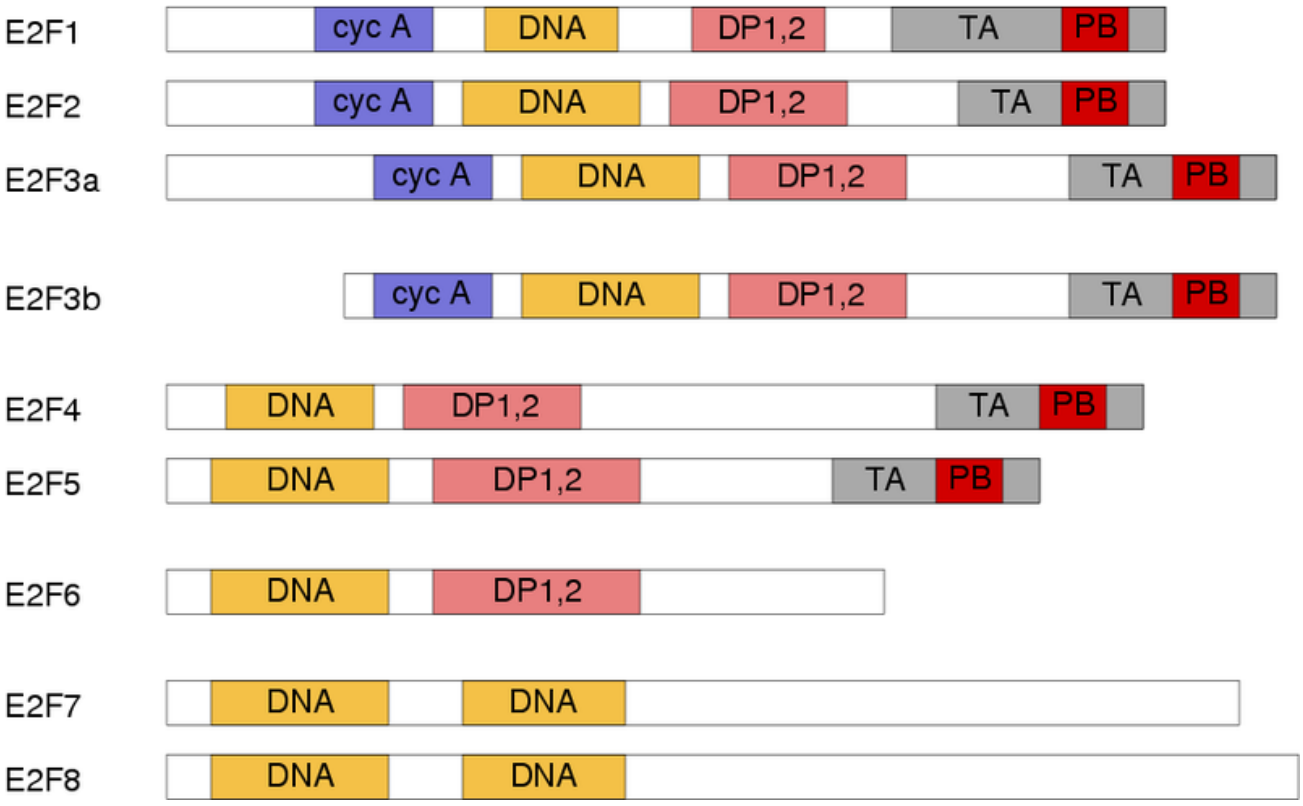

Figure 2.1. Comparison of the overall layouts of known E2f members in mammals

One problem arises in the comparative assays between the E2f groups between different organisms; in some of the less complex organisms, there are, at the very least, only two E2f groups used in the cell, one as an activator and one as a repressor. A possible reason for this discrepancy may lie in the specificity required for genes; in more complex organisms, it is possible that the different factors may specialize in regulating different genes, even among members of the same families. This may reflect an increase in complexity of cell cycle regulation in these higher organisms. To account for this problem in these analyses, the study will focus on one E2f member at a time; to ensure proper homology, we will try to use members from different organisms that have similar sequences or homology, as done through *in silico* analysis.

2.5 Methods Used for Analysis

To accomplish this in silico method of analysis, two major programs are to be used; the online ClustalW servers and the freely-downloadable application group PHYLIP 3.66.

Most of the known sequences used for this analysis are made freely available on the UC Santa Cruz Genome Viewer (http://genome.ucsc.edu/), and the NCBI PubMed website. This contains known E2f sequences for major model organisms as well as those of others whose sequences have been recently decoded. To make the most accurate analysis of the genetic relationships between the species, we will use the known proteome instead of genome sequences of the known E2f factors, due to codon redundancy issues that could skew the analysis and relative distances of the differing organisms. Also, proteomic sequences are those allowed in the ClustalW server for comparison purposes.

Once the correct sequences have been gathered and formatted correctly, they are then run through the ClustalW servers in at least two different fashions; for the first, comparisons are made, amino acid by amino acid, to establish the possible homology between any two factors. This is used primarily to establish which sequences of different organisms are suitable for usage in the analysis. For the comparisons to be properly made, the resulting output is analyzed in the .aln format.

Second, ClustalW also generates possible dendrograms based from the sequences selected for this purpose from different organisms. The trees are then stored as a text-only file in the Newick format for further analysis.

Finally, the PHYLIP 3.66 package is used in order to generate the correct phylogenetic trees for visual analysis from the text files yielded by the ClustalW server. Two of their applications, drawgram and drawtree, are used for this purpose to generate the diagrams that will be used throughout this thesis. However, one of the trees exceeds the size capacity for PHYLIP 3.66, so a special ClustalW server was used to generate it.

To solve the last question, another server was used to find the correct sequences. We used the NCBI BLAST website to enter in the sequences into the databases of particular organisms. In this case, we used the DNA-binding domain of human E2f1 for this purpose, searching for homologous sequences in other organisms. Some of the results of our search are given in Appendix B.

CHAPTER 3

LITERATURE SEARCH

3.1 Introduction

Previous studies referred to throughout this thesis were used as references to the functions of certain E2f factors. This was done to provide reasons and/or logical connections behind any predictions made through the analysis of phylogenetic patterns given through the known sequences of the E2f factors. While this thesis is based on these studies, only the most relevant literature will be briefly described in this chapter.

3.2 Literature Review

A brief comparison of the functions of these proteins may be appropriate for this. To garner this information, the files of each of the known factors as shown on the Entrez Gene servers were consulted [NCBI Entrez Gene Website]. Through these servers, the structural and functional homology of the E2f1-3 groups is clearly evident. Not only are these similar in domain structure, but the all groups also bind to the pRB domain in a cell-cycle dependent manner and all have additional cyclin binding domains not found in other sequences of E2fs. Furthermore, the Entrez Gene servers seem to indicate that the E2f factors are divided into four different groups; the activator ones with E2f1-3, the group composed of E2f4-5, E2f6, and the E2f7-8 group. Later analysis shown below confirms this division.

As for actual phylogenetic traits and differences among E2f factors, another paper may be useful for providing background information; in a paper published by Trimarchi and Lees [2002], it is indicated that the E2f groups may indeed have fallen into four categories, segregating into the different activator and repressor complexes commonly found in organisms, as shown in Figure 3.1 below.

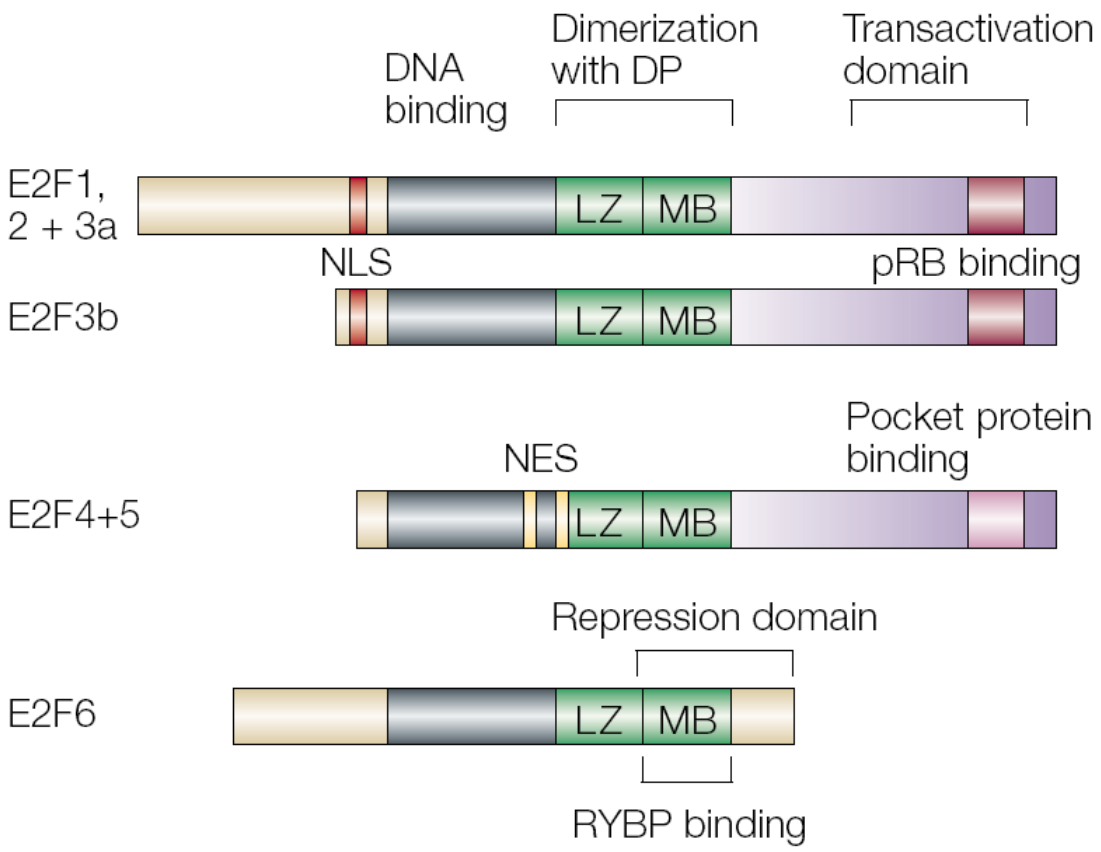

Figure 3.1 Division of the E2f proteins after Trimarchi and Lees [2002]

In the paper referred to above, it seems that the E2f factors have clear-cut roles, some acting as activators while others act as repressors. However, a later paper published by DeGregori and Johnson [2006] seems to indicate otherwise, citing from findings

pertaining to overexpression studies and characterizations of novel E2f factors. For instance, in some circumstances, if the E2f1 gene is somewhat overexpressed, then it can also de-express just as many genes as it activates them, and this action can also mediate apoptosis in cells. Although the functions of other E2f factors in the promotion of apoptosis is not as clearly defined, this trait shows that the boundaries of function could very well be considered as naïve and overly simplistic. Phylogenetic analysis of the E2f factors may reveal the possible “true grouping” of the factors; in combination with the functional analyses given, these may allow us to develop some hypothesis about the possible divergence of the groups.

CHAPTER 4

SEQUENCE ANALYSIS

4.1 Introduction

So far, the E2f protein sequences have been compiled for most of the major model organisms in genetic research and submitted to the public scientific community. Our primary sources for protein sequence data are PubMed and the UCSC Genome browser. Before going into the full phylogenetic analysis of the organisms based on these, some caveats need to be noted and addressed.

First, some organisms have significantly fewer than 8 E2f factors that have been characterized so far based on the sequence of the genome. For example, both *C. elegans* and *D. melanogaster* have only two E2f factors, the first for activating transcription and the second for transcriptional repression. The discrepancy was accounted for as shown below, using ClustalW analysis of homology. Table 4.1 below represents the different E2f sets that were aligned into the ClustalW servers to achieve the result (Italic abbreviations refer to the fact that the associated sequence has been predicted through in silico analysis):

Table 4.1. Tabulation of all the E2f sequences used that were found via searches from the UCSC and PubMed databases

SPECIES	**Align 1**	**Align 2**	**Align 3**	**Align 4**	**Align 5**	**Align 6**	**Align 7**	**Align 8**
Human	E2f1	E2f2	E2f3	E2f4	E2f5	E2f6	E2f7	E2f8
M. musculus	E2f1	E2f2	E2f3	E2f4	E2f5	E2f6	E2f7	E2f8
G. gallus	E2f1	---	*E2f3*		E2f5		*E2f7*	*E2f8*
Danio rerio	*E2f1*	---		E2f4			E2f7	
D. melanogaster	dE2f1	dE2f1	dE2f1	dE2f2	dE2f2	dE2f2	dE2f2	dE2f2
Chlamydomonas	E2f1	---	---	---	---	---	---	---
Arabidopsis	E2f1	E2f2	E2f3	E2f4	E2f5			
C. elegans	efl-1	efl-1	efl-1	efl-2	efl-2	efl-2	efl-2	efl-2

4.2 Analysis

To start, it is interesting to note the intense conservation of some of the protein sequences among organisms. Judging from the .aln file also resulting from preliminary ClustalW analysis of select model E2f1s, only amino acids 218-282 of the human E2f1 sequence are conserved among other organisms. Furthermore, these correspond to the DBD binding domain of the E2f protein; the largely conserved region almost exactly aligns with that very region in the sequence, as shown in Figure 4.1.

```
HSE2f1          HPGKGVKSPGEKS**RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQ-KRRIYDITN**
ChlamE2f1       CAAGSPGSHTGGCRYDSSLGMLTKKFLNLINTARDGILDLNQAAETLKVQ-KRRIYDITN
ArabidoE2f1     GSPGNNFAQAGTCRYDSSLGLLTKKFINLIKQAEDGILDLNKAADTLEVQ-KRRIYDITN
MusE2f1         HPGKGVKSPGEKSRYETSLNLTTKRFLELLSRSADGVVDLNWAAEVLKVQ-KRRIYDITN
GalE2f1         IPGRGAKSPGEKSRYETSLNLTTKRFLELLSQSPDGVVDLNWAAEVLKVQ-KRRIYDITN
DmelE2f1        ASVASSSSSGDRNRADTSLGILTKKFVDLLQESPDGVVDLNEASNRLHVQ-KRRIYDITN
CEefl-1         -EDEDLDQPQMGTRADKSLGLLAKRFIRMIQYSPYGRCDLNTAAEALNVRQKRRIYDITN
DanioE2f1       APPRVPKLAVEKSRYDTSLNLTTKRFLDLLAQSPDGVVDLNWASQVLDVQ-KRRIYDITN
                             * :.**.: :*:*: ::  :  *  *** *:: *.*: *********

HSE2f1          **VLEGIQLIAKKSKNHIQWL**GSHTTVG-----------VGGRLEGLTQDLRQLQESEQQL
ChlamE2f1       VLEGVGLIEKKSKNNIRWKGAGDGGRGG---------DADPDLDRLRSDMSKLD--EREL
ArabidoE2f1     VLEGIGLIEKTLKNRIQWKGLDVSKPG----------ETIESIANLQDEVQNLAAEEARL
MusE2f1         VLEGIQLIAKKSKNHIQWLGSHTMVG-----------IGKRLEGLTQDLQQLQESEQQL
GalE2f1         VLEGIQLITKKSKNNIQWLGSQVAAG-----------ASSRQRLLEKELRDLQAAERQL
DmelE2f1        VLEGINILEKKSKNNIQWRCGQSMVS-----------Q-ERSRHIEADSLRLEQQENEL
CEefl-1         VLEGIGLIEKRSKNMIQWKGGDFMLNVKEGKRQSATTEEEDRMEQLKAEIEQLNKEEELI
DanioE2f1       VLEGIHLISKKSKNNIQWLGNRIDGA-----------SLARFQELQKEVSELTEAEEKL
                ****: :: *  ** *:*                          :  :   *   *  :
```

Figure 4.1 Comparison of E2f1 sequences in eight of the model organisms used.

Taking advantage of the general homology of this region among the different E2f factors, we have also done BLAST searches using the DNA-binding domain sequence of human E2f1 (highlighted in bold in the above figure) to look for potential E2fs in other model organisms, particularly in Danio rerio and Arabidopsis thaliana, thereby addressing one of the questions given in the beginning of this thesis.

By inputting the relevant E2f sequences, we used ClustalW and PHYLIP 3.66 to analyze the resulting dendrogram. We have been able to yield the following result based on the multiple E2f1 sequences as given in Figure 4.2 below:

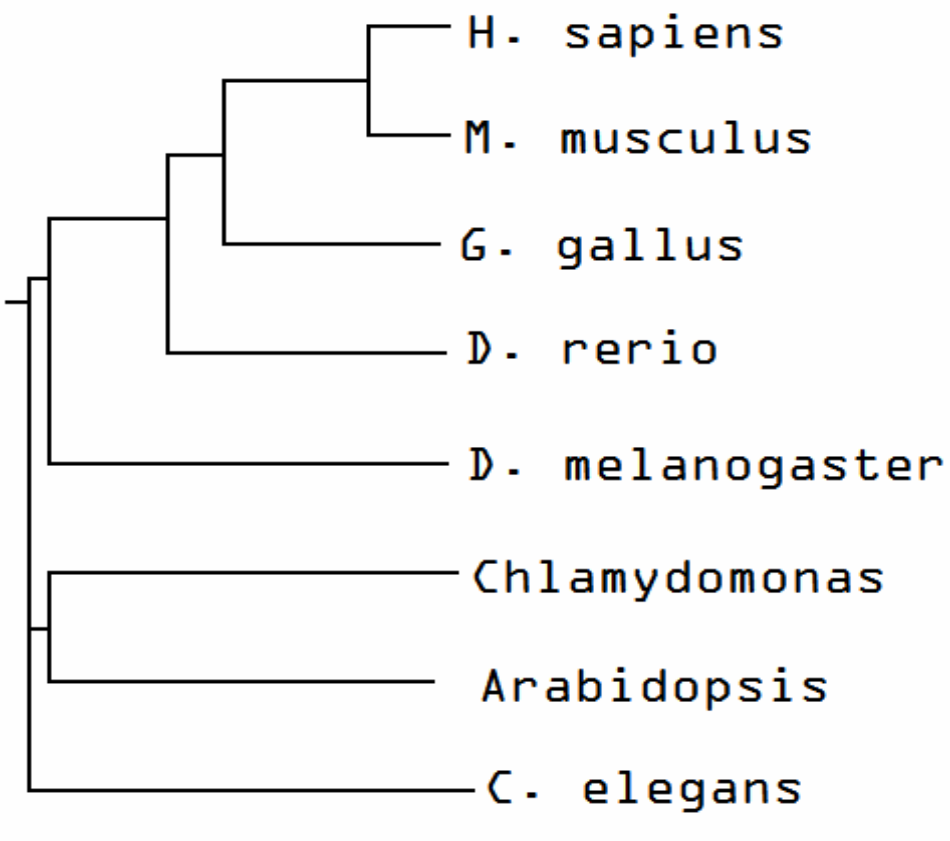

Figure 4.2 Phylogenetic tree for E2f1 analysis in different organisms

Upon further analysis of the resulting dendrogram and phylogenetic tree from ClustalW, an interesting observation can be pointed out. In the branch, C. elegans seems to be the earliest to diverge, even before the rise of the two organisms Chlamydomonas and Arabidopsis, the former being a protist and the latter being a plant. Usage of the other efl-2 sequence of C. elegans for this (not shown) also yields

this pattern. Other than this, the results seem to correlate with the consensus phylogeny. Why this discrepancy is present in the resulting tree is still under investigation, and calls into question the efficacy of using E2f sequences in arranging phylogeny of organisms.

Perhaps the most fascinating find is the phylogeny hidden between the different mammalian groups, which could in fact give a certain organization as to how these proteins, not just the organisms they were contained in, actually evolved over time. Such comparison will be done on the two different known groups of E2f factors, the activator factors (E2f1-3) and the repressor factors (E2f4-8).

4.3 Comprehensive E2f analysis

The next step in research is the actual evolutionary patterns surrounding the E2f factors. We know that the E2f factors have currently diverged into the eight groups discovered thus far in mammalian cells. The question remains, however: as there are discrepancies between the numbers of E2f groups among organisms of differing complexities, does this indicate a possible evolutionary relationship between the two groups, and if so, in what ways? To accomplish this task, several different organismal E2f sequences were garnered throughout the scientific community to make the full comparison and sequence analysis of the organisms and pathways.

In preparation for ClustalW analysis of the compounds, certain postulates were made for testing through this method. One postulate made organization clearer for the purposes of this thesis: no matter the number of factors given in a particular organism or its overall complexity, there are almost always at least two factors that have been characterized, the promoter proteins and the repressor proteins. It can thus be speculated that the E2f promoter proteins and the E2f repressor proteins may have evolved separately into the two different loci that are known. Therefore, for the initial analysis, the two different groups of protein will be analyzed separately, but then they will be exposed to full analysis through the ClustalW servers.

Through analysis based on activator mammalian E2f sequences, one tentative result to be given is shown in the following tree (Figure 4.3), which contains E2f1-3 sequences from known organisms.

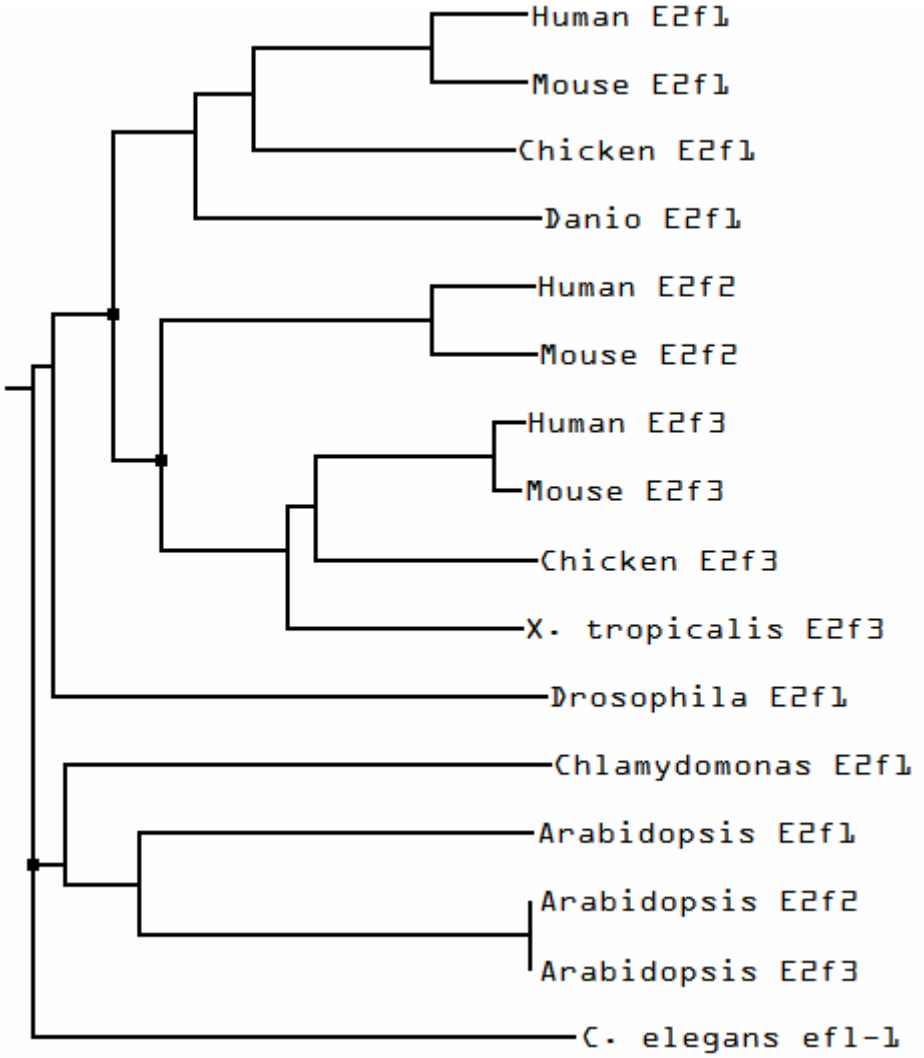

Figure 4.3 Comparison between the E2f1-3 groups of organisms shown in Table 4.1.

The results of this tree are substantial in determining predictions as to the evolutionary break points of the activator E2f factors. Especially of importance are the possible patterns of evolutionary development that were taken with the different groups throughout organismal development. The highlighted nodes (filled in to facilitate deduction) seem to indicate that the E2f factors giving rise to more complex organisms like humans and/or mice diverged first into the precursor group for mammalian E2f1-3,

with the other group diverging into the factors present in plants such as Arabidopsis and Chlamydomonas. Once again, it is important to note that C. elegans forms yet a third group that seems out of place in this tree; considering this has occurred with other sequence comparisons, it is possible that, even though C. elegans is technically considered as an animal, this inconsistency may apparently be a result of convergent evolution in function.

Among the groups that have been discussed, it can be shown through the same analysis that E2f1 seems to have diverged first from the tree, apparently during the same time that amphibians diverged from fish. E2f2 and E2f3 apparently diverged soon afterwards, before amphibians evolved further. Although at this stage it is not clear as to how this has happened, certain analyses of the functions of the different factors, such as those on their possible differences in function or redundancies in cell transcription regulation, may be important to provide a hypothesis as to the patterns of evolutionary divergence in the E2f sequences.

The other set to be studied, however, does not submit as easily to phylogenetic analysis. Given the size of the group in question, this may lead to some complexity on the part of evolutionary relationships. The results of ClustalW analysis are shown below in Figure 4.4:

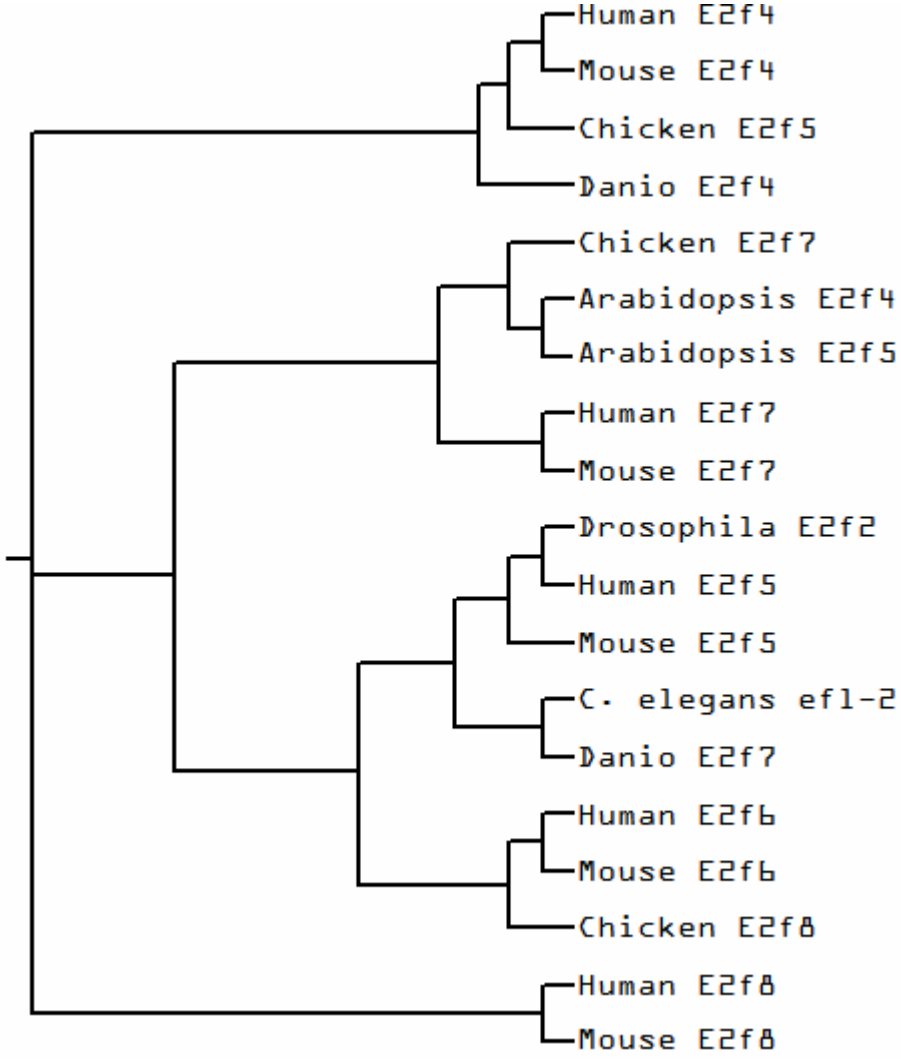

Figure 4.4 A comparison of different E2f repressor groups in different model organisms

In this scenario, it seems apparent that, at least for the repressor proteins, each of the E2f groups for humans and mice seem to arise into separate groups; two of the more primitive groups include those for E2f4 and especially E2f8 while the others seem to have diverged in later periods. Of these, E2f5 and E2f6 seem to have diverged last, each factor diverging into their respective human and mouse homologs.

At first glance, this may indicate that the groups may indeed have evolved from either two ancestral genes, or even from one gene that diverged into two primitive E2f genes. However, this assumption is shown to be unsatisfactory. The table shown only displays canonical E2f genes taken from biological literature, and thus is not completely representative of this evolutionary pathway. As one of our objectives is to yield a comprehensive evolutionary pathway for this experiment, one of the most definitive ways to do so would be to analyze the sequences all at once. Therefore, such analysis will be summarized below in Figures 4.5 and 4.6 below:

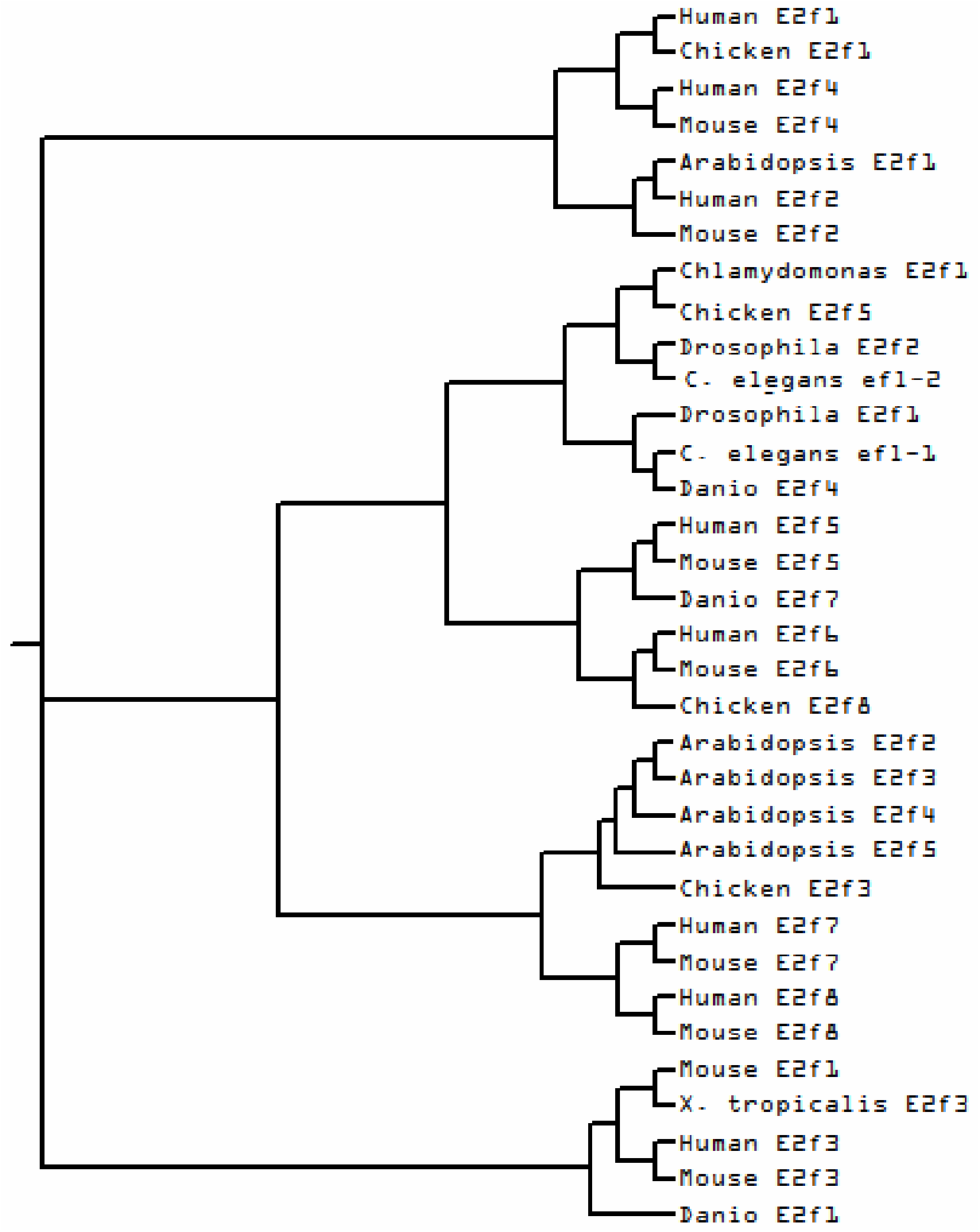

Figure 4.5 A comparison between all the known E2fs considered in Table 1

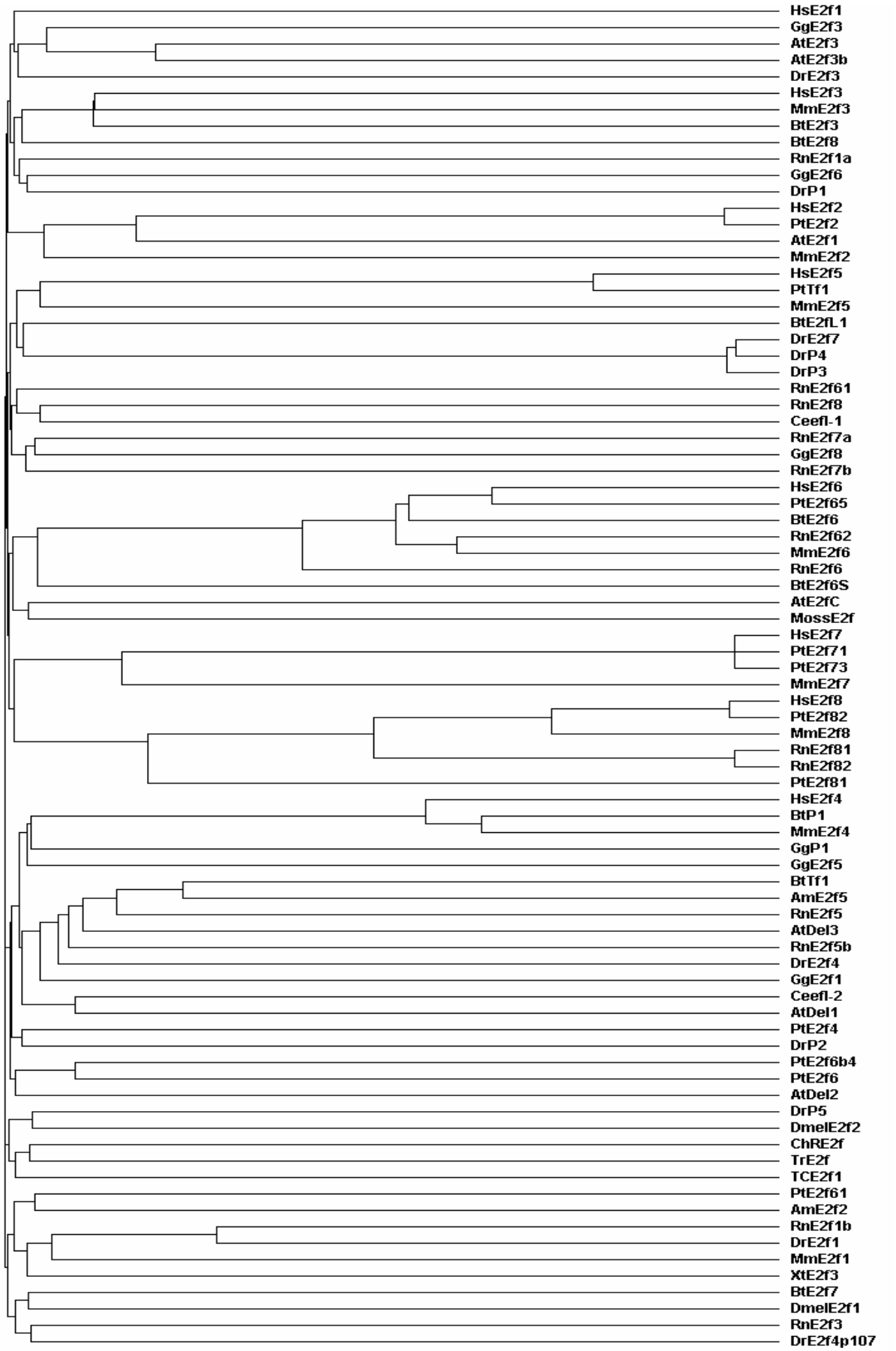

Figure 4.6 All E2fs considered in Table 4.1 and found through BLAST analysis

The first tree is derived from adding all the factors shown in the trees detailing the activator and the repressor factors. However, the second tree includes alternate factors not derived directly from searches of the E2f factor in organisms. These were found using BLAST searches of genomes of different organisms with the human E2f1 DNA-binding domain in other organisms, including those omitted from previous trees. Due to the sheer size of the sequence information, however, the second ClustalW server had to be used; thereby abbreviating the groups as shown in Table 4.2 below:

Table 4.2 Explanation of symbols used in Figure 4.6

Species Name	**Abbreviation**
Homo sapiens (human)	Hs
Pan troglodytes (chimpanzee)	Pt
Bos taurus (cow)	Bt
Rattus norvegicus (rat)	Rn
Mus musculus (mouse)	Mm
Gallus gallus (wild fowl)	Gg
Danio rerio (zebrafish)	Dr
Drosophila melanogaster	Dmel
Caenorhabditis elegans	Ce
Chenopodium rubrum (red goosefoot)	ChR
Arabidopsis thaliana (wild mustard)	At
Apis mellifera	Am
Thlaspi caerulescens	TC
Xenopus tropicalis	Xt
Physcomitrella patens	Moss
Triticum sp. (wheat)	Tr

The tree in Figure 4.6 seems to uncover new details about E2f inter-relationships in this analysis. The resulting tree from this analysis turns out to be a form of hybrid design of phylogeny that still somewhat keeps the activator/repressor dichotomy, though the individual groups have been heavily reorganized throughout the tree. The pattern of reorganization of mammalian E2f factors, for instance, is shown below in Figure 4.7:

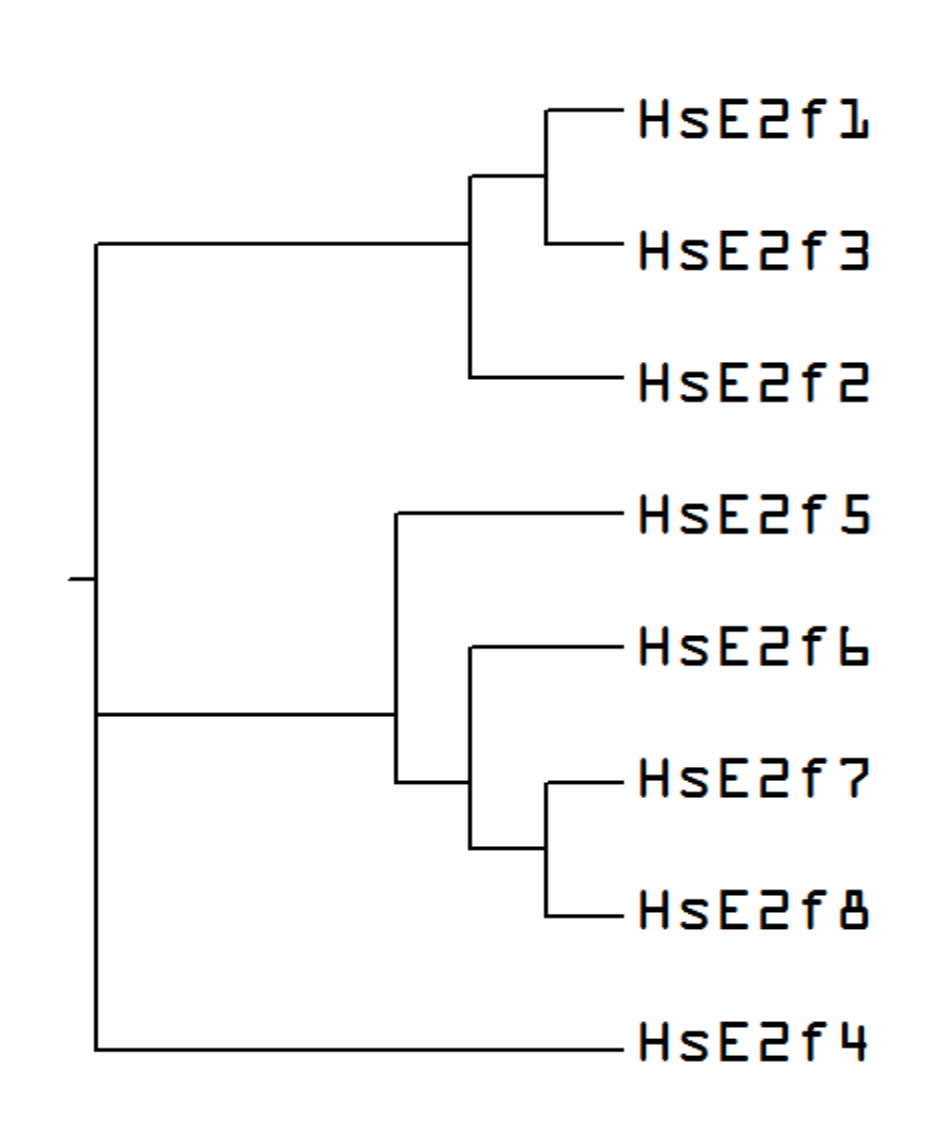

Figure 4.7 Organization of E2f factors as simplified from the tree shown in Figure 4.6

4.4 Search for possible E2f factors

It should be noted that to create a full compilation of the E2f analysis, further techniques were used; one of the main objectives of the ideal within this honors thesis is not only to discern possible phylogenies from known characterized sequences of E2f factors in various model organisms, but also to find potential new E2f factors in the model organisms analyzed.

In order to fully analyze the possibility of continued E2fs, we decided to find the defining characteristic of most E2fs first. As stated in the above stage of analysis, we discovered that the DNA-binding domain would be the best criterion to use for this analysis, because it is one of the only conserved domains in all the compared E2fs, so would thereby be an excellent distinguishing feature of the factor classes. The results of these analyses as were shown on the NCBI website will be given in the appendix.

One of the major starting points for search would be using the Danio rerio database to discern possible E2f factors and their potential percent homologies using this sequence, both using the DNA-binding domain and the factor as a whole for this sequential analysis. Searches on the NCBI BLAST databases revealed not only the canonical E2f factors that were used in Table 4.1, but also other potential proteins in the fish, some with lower homologies to the DBD than others. The same was true for searches in the Arabidopsis thaliana genome. However, when searching the genomes for less complex metazoans such as Drosophila melanogaster and Apis mellifera (honeybee), only two results each were discovered; no new results were present in humans or mice either, but this is probably due to the fact that all the known E2f genes

in these organisms had already been characterized. In other mammals such as Pan troglodytes and *Bos taurus*, some potential E2f-like factors and proteins were found; however, the fact that they have not been canonically represented as E2fs may be due to the fact that they are still under investigation within the scientific community.

The results of BLAST search are listed in Appendix B, but the following proteins that are shown in the above tree are purported E2f factors in organisms: in Danio rerio, we have DrP1, DrP2, DrP3, DrP4, and DrP5. From Pan troglodytes, PtTf1, while in Gallus gallus, GgP1, with BtP1, BtTf1, and BtE2fL1. The NCBI identity numbers of these and others are given in Appendix B.

CHAPTER 5

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Summary

This thesis introduces methods to organize the E2f factors, compiling a comprehensive phylogeny of them to determine how they could possibly have diverged from one another. There were three tasks to do so: first, to gather sequences and to compare their homologies to yield the set for a comprehensive analysis; second, to fully analyze the resulting phylogenies to be used; and third, to search for novel putative E2f factors through database analysis.

In order to do this, bioinformatics software was used to compute the primary alignments necessary to establish any conclusions. The software used for this *in silico* analysis primarily laid within the ClustalW alignment servers. This software, with its large sequence capacity, is employed to gather alignment data of the various sequences. All the Newick dendrograms used in this program were made by the ClustalW servers and then visually processed with the PHYLIP 3.66 application.

Another primary venue of analysis was done via BLAST analysis; the server used for this purpose was the one provided directly by NCBI. This was done in the interest of searching for any new and putative E2f factors, a task that lay as one of the objectives of this thesis.

Finally, to set the backdrop for our analysis and any conclusions, we looked up some papers about the phylogeny of the E2f factors. The results of this search indicate

distinctions between the two groups, although it turns out that the boundaries are not as clear as was initially thought.

5.2 Conclusions

The data indicated from the phylogenetic trees seem to indicate a novel phylogeny among E2f sequences in organisms; originally, it was expected that the order would evenly split into two separate groups; the cellular activator sequences and the repressor factors. However, even preliminary trees seem to somewhat weaken this theory. True, some homology is maintained between the different activator/repressor classes, but in the preliminary tree, not two, but three groups are maintained. This find in phylogeny supports a theory established by DeGregori and Johnson from their work in 2006, which indicates that some of the activator groups may function in transcriptional repression as well.

Further analysis of other putative factors revealed through BLAST analysis seems to alleviate this discrepancy, however. It is noticeable that the groups E2f1-3 tend to segregate into one branch of the tree. However, group similarities between the factors are no longer distinct; while each human factor still lies within its own group, other similarly numbered/related factors actually have homologies with factors from other organisms instead. Furthermore, the order of the groups that are more closely related to one another in the tree are different; although the groups composed of 1-3 and 7-8 are still related to one another quite closely, the other groups cluster in different locations throughout the tree. Assuming the tree is indeed accurate and comprehensive

for our purposes, this suggests that certain group members may have come closer to one another via convergent evolution, possibly changing from other, closely related factors.

Finally, the BLAST search does indeed reveal new potential E2f factors, based on their similarity with the potential DNA-binding domain. Overall homology in many of the designated factors as revealed through this search is not very extensive in some of the lower organisms like Arabidopsis (40-50% homology in the DNA-binding domain), while in other, more complex organisms, we achieve about 60 to 70% homology; some of the domains, it must be noted, are almost identical to standard DBDs, with around 90% homology, which would direct our interest towards exploring the function of these proteins. However, it must be said that searches in humans, mice, and Drosophila did not yield any additional sequences, possibly because the genes themselves had already been characterized by now. Further investigation is needed to completely discern the function of these putative protein sequences.

5.3 Recommendations

This thesis introduces a first attempt to better organize the different E2f factors in a phylogenetically logical manner. Further studies are recommended as follows:

1. This study can be expanded to track the E2f factors in additional plants, microbes, and fungi as well as the metazoans considered thus far. Although it may be true that these organisms use less advanced methods of cell regulation, indeed they may not even have E2f factors, it may be helpful to see what possible evolutionary roots may

have occurred in the development of the primary E2f factors in the later metazoan organisms. Earlier drafts of the tree were significantly different due to the omission of some key sequences. It must also be noted that the organisms used were the ones with their genomes completely sequenced.

2. In the search for more factors, other media should also be used in future assays. In this study, most of the factors were found using BLAST searches on particular databases with the DNA-binding domain of human E2f1. Although we suspect that this portion is the main distinguishing feature of any E2f factor due to its intense conservation among different E2f factors, both within humans and among other organisms, results might also be obtained if other domains were also used, or even if multiple domains were inputted into the ClustalW server.
3. Finally, for the purposes of comparison, percent homologies between the different factors should also be considered. If the divergence limits are not enough, then the actual phylogenetic changes may in fact be caused by other phenomena, including genetic drift, selective pressure, and cell cycle regulation requirements and multicellular complexity.

List of References:

1. ClustalW Server, http://www.ch.embnet.org/software/ClustalW.html, latest date accessed May 20, 2007

2. ClustalW Server, http://www.ebi.ac.uk/clustalw/, latest date accessed May 21, 2007

3. DeGregori, J., and Johnson, D., *Distinct and Overlapping Roles for E2f Family Members in Transcription, Proliferation, and Apoptosis.* Current Molecular Medicine, **6**, 739-748 (2006)

4. Logan, N., Graham, A., Zhao, X., Fisher, R. Maiti, B., Leone, G., and La Thangue, N. *E2F-8: an E2F family member with a similar organization of DNA-binding domains to E2F-7*. *Oncogene* (2005) **24,** 5000–5004

5. NCBI Entrez Gene, latest date accessed May 21, 2007 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene,

6. NCBI PubMed BLAST Server, http://www.ncbi.nlm.nih.gov/BLAST/, latest date accessed May 21, 2007

7. Qiao, H., Stefano, L. et al. 2006. *Human TFDP3, a Novel DP Protein, Inhibits DNA Binding and Transactivation by E2F*. The Journal of Biological Chemistry 282(1):454-466.

8. Trimarchi, J. M., and Lees, J. A. (2002) *Sibiling Rivalry in the E2f Family*. Nature Review Molecular Cell Biology. **3**, 11-20

9. UCSC Genome Browser, UC Santa Cruz, http://genome.ucsc.edu/, latest date accessed April 2007

Appendix A: List of Abbreviations:

Species Name	Abbreviation
Homo sapiens (human)	Hs
Pan troglodytes (chimpanzee)	Pt
Mus musculus (mouse)	Mm
Gallus gallus (wild fowl)	Gg
Danio rerio (zebrafish)	Dr
Drosophila melanogaster	Dmel
Caenorhabditis elegans	Ce
Chenopodium rubrum (red goosefoot)	ChR
Arabidopsis thaliana (wild mustard)	At
Apis mellifera	Am
Thlaspi caerulescens	TC
Xenopus tropicalis	Xt
Physcomitrella patens	Moss
Triticum sp. (wheat)	Tr
Putative Protein	P
Transcription factor-like	Tf
E2f-like	E2fL
E2f Splice	E2fS

Appendix B: BLAST searches shown throughout this thesis

Arabidopsis thaliana

```
>ref|NP_175222.1| UniGene infoGene info E2FC (ARABIDOPSIS HOMOLOG OF E2F C);
transcription factor [Arabidopsis thaliana] Length=396 (AtE2fC)

 Score = 88.6 bits (218),  Expect = 7e-19
 Identities = 41/64 (64%), Positives = 53/64 (82%), Gaps = 0/64 (0%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY++SL L TK+F++L+  + DG +DLN+ A VL+VQKRRIYDITNVLEGI LI K +KN
Sbjct  155  RYDSSLGLLTKKFVKLIQEAEDGTLDLNYCAVVLEVQKRRIYDITNVLEGIGLIEKTTKN  214

Query  61   HIQW  64
            HI+W
Sbjct  215  HIRW  218


>ref|NP_568413.1| UniGene infoGene info E2F1; transcription factor [Arabidopsis thaliana]
 ref|NP_001031921.2| Gene info E2F1; transcription factor [Arabidopsis thaliana]
Length=469 (AtE2f1)

 Score = 86.7 bits (213),  Expect = 3e-18
 Identities = 41/64 (64%), Positives = 51/64 (79%), Gaps = 0/64 (0%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY++SL L TK+F+ L+  + DG++DLN AA+ L+VQKRRIYDITNVLEGI LI K  KN
Sbjct  129  RYDSSLGLLTKKFINLIKQAEDGILDLNKAADTLEVQKRRIYDITNVLEGIGLIEKTLKN  188

Query  61   HIQW  64
             IQW
Sbjct  189  RIQW  192


>ref|NP_565831.3| Gene info E2F3 (E2F TRANSCRIPTION FACTOR-3); transcription factor
[Arabidopsis thaliana] Length=483 (AtE2f3)

 Score = 84.7 bits (208),  Expect = 1e-17
 Identities = 41/64 (64%), Positives = 50/64 (78%), Gaps = 0/64 (0%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY++SL L TK+F+ L+  + DG++DLN AAE L+VQKRRIYDITNVLEGI LI K  KN
Sbjct  167  RYDSSLGLLTKKFVNLIKQAKDGMLDLNKAAETLEVQKRRIYDITNVLEGIDLIEKPFKN  226

Query  61   HIQW  64
             I W
Sbjct  227  RILW  230


>ref|NP_973610.1| Gene info E2F3 (E2F TRANSCRIPTION FACTOR-3) [Arabidopsis thaliana]
Length=514 (AtE2f3b)

 Score = 76.3 bits (186),  Expect = 4e-15
 Identities = 37/58 (63%), Positives = 45/58 (77%), Gaps = 0/58 (0%)

Query  7    NLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKNHIQW  64
            +L TK+F+ L+  + DG++DLN AAE L+VQKRRIYDITNVLEGI LI K  KN I W
Sbjct  219  SLLTKKFVNLIKQAKDGMLDLNKAAETLEVQKRRIYDITNVLEGIDLIEKPFKNRILW  276
```

>ref|NP_186782.2| Gene info DEL3 (DP-E2F-like 3); transcription factor [Arabidopsis thaliana] Length=354 (**AtDel3**)

```
 Score = 52.8 bits (125),  Expect = 4e-08
 Identities = 25/64 (39%), Positives = 38/64 (59%), Gaps = 0/64 (0%)

Query  1   RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
           R E SL +    FL L +     ++ L+ AA  L V++RRIYD+ N+LE I ++A++ KN
Sbjct  21  RKEKSLGVLVSNFLRLYNRDDVDLIGLDDAAGQLGVERRRIYDVVNILESIGIVARRGKN  80

Query  61  HIQW  64
              W
Sbjct  81  QYSW  84


>ref|NP_851012.1| UniGene infoGene info DEL1 (DP-E2F-like 1); transcription factor [Arabidopsis thaliana] Length=379 (AtDel1)

 Score = 51.2 bits (121),  Expect = 1e-07
 Identities = 26/64 (40%), Positives = 38/64 (59%), Gaps = 0/64 (0%)

Query  1   RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
           R + SL L    FL L +     +V L+ AA  L V++RRIYDI NVLE + ++ +++KN
Sbjct  34  RKQKSLGLLCTNFLALYNREGIEMVGLDDAASKLGVERRRIYDIVNVLESVGVLTRRAKN  93

Query  61  HIQW  64
              W
Sbjct  94  QYTW  97


 Score = 35.8 bits (81),  Expect = 0.005
 Identities = 24/67 (35%), Positives = 37/67 (55%), Gaps = 11/67 (16%)

Query  1    RYETSLNLTTKRFLEL-LSHSADGVVDLNWAAEVL----------KVQKRRIYDITNVLE  49
            R E SL L T+ F++L +   A  ++ L+ AA++L          + + RR+YDI NVL
Sbjct  169  RREKSLGLLTQNFIKLFICSEAIRIISLDDAAKLLLGDAHNTSIMRTKVRRLYDIANVLS  228

Query  50   GIQLIAK  56
             + LI K
Sbjct  229  SMNLIEK  235


>ref|NP_197000.1| UniGene infoGene info DEL2/E2FD/E2L1 (DP-E2F-LIKE 2); DNA binding / transcription factor [Arabidopsis thaliana] Length=359 (AtDel2)

 Score = 49.3 bits (116),  Expect = 5e-07
 Identities = 25/64 (39%), Positives = 36/64 (56%), Gaps = 0/64 (0%)

Query  1   RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
           R + SL +    FL L +     +  L+ AA  L V++RRIYD+ N+LE I L+A+  KN
Sbjct  13  RKDKSLGVLVANFLTLYNRPDVDLFGLDDAAAKLGVERRRIYDVVNILESIGLVARSGKN  72

Query  61  HIQW  64
              W
Sbjct  73  QYSW  76
```

Danio rerio:

```
>ref|XP_695874.2|  PREDICTED: similar to E2F-1 transcription factor [Danio rerio]
Length=431 (DrE2f1)

 Score =  112 bits (280),  Expect = 5e-26
 Identities = 54/65 (83%), Positives = 61/65 (93%), Gaps = 0/65 (0%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY+TSLNLTTKRFL+LL+ S DGVVDLNWA++VL VQKRRIYDITNVLEGI LI+KKSKN
Sbjct  125  RYDTSLNLTTKRFLDLLAQSPDGVVDLNWASQVLDVQKRRIYDITNVLEGIHLISKKSKN  184

Query  61   HIQWL  65
            +IQWL
Sbjct  185  NIQWL  189


>ref|XP_697294.1|  PREDICTED: hypothetical protein [Danio rerio]
Length=438 (DrP1)

 Score =  108 bits (270),  Expect = 8e-25
 Identities = 52/65 (80%), Positives = 60/65 (92%), Gaps = 0/65 (0%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY+TSL L TK+F+ LLS SADGV+DLNWA+EVL+VQKRRIYDITNVLEG+QLI KKSKN
Sbjct  133  RYDTSLGLLTKKFVGLLSESADGVLDLNWASEVLEVQKRRIYDITNVLEGVQLIRKKSKN  192

Query  61   HIQWL  65
            +IQWL
Sbjct  193  NIQWL  197


>ref|XP_688126.2|  PREDICTED: similar to E2F transcription factor 3 [Danio rerio]
Length=409 (DrE2f3)

 Score = 95.5 bits (236),  Expect = 7e-21
 Identities = 46/65 (70%), Positives = 56/65 (86%), Gaps = 0/65 (0%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY+TSL   TK+F +LL+ S+DGV+DLN AA VL VQKRR+YDITNVLEG++LI KKSKN
Sbjct  130  RYDTSLGFLTKKFCQLLAQSSDGVLDLNKAAIVLNVQKRRLYDITNVLEGVRLIKKKSKN  189

Query  61   HIQWL  65
            +IQWL
Sbjct  190  NIQWL  194


>ref|NP_998597.1|  E2F transcription factor 4 [Danio rerio] Length=393 (DrE2f3)

 Score = 79.7 bits (195),  Expect = 4e-16
 Identities = 43/65 (66%), Positives = 49/65 (75%), Gaps = 1/65 (1%)

Query  1   RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKV-QKRRIYDITNVLEGIQLIAKKSK  59
           R+E SL L T +F+ LL  + DGV+DL  AA+ L V QKRRIYDITNVLEGI LI KKSK
Sbjct  12  RHEKSLGLLTTKFVTLLQEAKDGVLDLKAAADTLAVRQKRRIYDITNVLEGIGLIEKKSK  71

Query  60  NHIQW  64
           N IQW
Sbjct  72  NSIQW  76
```

>ref|NP_991178.1| E2F transcription factor 4, p107/p130-binding [Danio rerio]
Length=143 (**DrE2f4p107**)

 Score = 77.0 bits (188), Expect = 3e-15
 Identities = 42/65 (**64%**), Positives = 49/65 (**75%**), Gaps = 1/65 (1%)

```
Query  1   RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKV-QKRRIYDITNVLEGIQLIAKKSK  59
           R+E SL L T +F+ LL  + DGV+DL  AA+ L V QKRRIYDITNVLEGI LI KK+K
Sbjct  19  RHEKSLGLLTVKFVTLLQEAKDGVLDLKVAADSLAVKQKRRIYDITNVLEGIGLIEKKTK  78

Query  60  NHIQW  64
           N IQW
Sbjct  79  NTIQW  83
```

>ref|NP_001025315.1| hypothetical protein LOC560495 [Danio rerio]
Length=405

 Score = 71.2 bits (173), Expect = 1e-13
 Identities = 34/65 (**52%**), Positives = 47/65 (**72%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            R E +L   TKRF++LL+ + +GV+DLN  +  L  +KRR+YDIT+VL GI L+ K SKN
Sbjct  182  RSEVALGQLTKRFMQLLNAAPEGVLDLNEVSRKLGARKRRVYDITSVLAGIHLLKKTSKN  241

Query  61   HIQWL  65
             IQW+
Sbjct  242  KIQWM  246
```

>ref|XP_001341422.1| PREDICTED: hypothetical protein [Danio rerio] Length=432

 Score = 48.5 bits (114), Expect = 1e-06
 Identities = 27/68 (**39%**), Positives = 42/68 (**61%**), Gaps = 4/68 (5%)

```
Query  1    RYETSLNLTTKRFLELLS---HSADGV-VDLNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
            R + SL L  ++FL L      S++ + + L+  A  L V++RRIYDI NVLE + L+++
Sbjct  147  RKQKSLGLLCQKFLALYPDYPESSESINISLDEVATCLGVERRRIYDIVNVLESLMLVSR  206

Query  57   KSKNHIQW  64
            K+KN   W
Sbjct  207  KAKNMYVW  214
```

 Score = 35.4 bits (80), Expect = 0.008
 Identities = 23/69 (**33%**), Positives = 35/69 (**50%**), Gaps = 13/69 (18%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVL-------------KVQKRRIYDITNV  47
            R + SL + +++F+ L   S    V L+ AA++L             K + RR+YDI NV
Sbjct  264  RKDKSLRIMSQKFVMLFLVSKTQTVTLDMAAKILIEEGQEESYDSKYKTKVRRLYDIANV  323

Query  48   LEGIQLIAK  56
            L  + LI K
Sbjct  324  LTSLNLIKK  332
```

>ref|NP_001038612.1| hypothetical protein LOC567941 [Danio rerio] Length=704

 Score = 48.5 bits (114), Expect = 1e-06
 Identities = 27/68 (**39%**), Positives = 42/68 (**61%**), Gaps = 4/68 (5%)

```
Query  1    RYETSLNLTTKRFLELLS---HSADGV-VDLNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
```

```
            R + SL L  ++FL L      S++ + + L+  A  L V++RRIYDI NVLE + L+++
Sbjct  147  RKQKSLGLLCQKFLALYPDYPESSESINISLDEVATCLGVERRRIYDIVNVLESLMLVSR  206

Query  57   KSKNHIQW  64
            K+KN   W
Sbjct  207  KAKNMYVW  214


 Score = 35.4 bits (80),  Expect = 0.008
 Identities = 23/69 (33%), Positives = 35/69 (50%), Gaps = 13/69 (18%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVL-------------KVQKRRIYDITNV  47
            R + SL + +++F+ L   S    V L+ AA++L             K + RR+YDI NV
Sbjct  264  RKDKSLRIMSQKFVMLFLVSKTQTVTLDMAAKILIEEGQEESYDSKYKTKVRRLYDIANV  323

Query  48   LEGIQLIAK  56
            L  + LI K
Sbjct  324  LTSLNLIKK  332


>ref|XP_694311.2|  PREDICTED: hypothetical protein [Danio rerio]
Length=866

 Score = 42.0 bits (97),  Expect = 9e-05
 Identities = 23/68 (33%), Positives = 38/68 (55%), Gaps = 4/68 (5%)

Query  1    RYETSLNLTTKRFLELLSH----SADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
            R + SL L   +FL    +    + +  + L+  A  L V++RRIYDI NVLE + ++++
Sbjct  75   RKDKSLGLLCYKFLARYPNYPNPALNNGISLDDVAAELHVERRRIYDIMNVLESLNMVSR  134

Query  57   KSKNHIQW  64
             +KN   W
Sbjct  135  LAKNRYTW  142


 Score = 37.4 bits (85),  Expect = 0.002
 Identities = 24/70 (34%), Positives = 38/70 (54%), Gaps = 14/70 (20%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVL--------------KVQKRRIYDITN  46
            R + SL + +++F+ L   S+  VV L+ AA++L              K + RR+YDI N
Sbjct  222  RKDKSLRVMSQKFVMLFLVSSPPVVSLDVAAKILIGEDHVVDQDKNKFKTKIRRLYDIAN  281

Query  47   VLEGIQLIAK  56
            VL  ++LI K
Sbjct  282  VLSSLELIKK  291
```

Bos taurus

>ref|XP_615437.3| PREDICTED: similar to transcription factor E2F like protein [Bos taurus] Length=579 (**BtE2FL1**)

Score = 126 bits (316), Expect = 3e-30
Identities = 63/65 (**96%**), Positives = 63/65 (**96%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RYETSLNLTTKRFLELLS SADGVVDLNWAAEVLKVQKRRIYDITNVLEGI LIAKKSKN
Sbjct  269  RYETSLNLTTKRFLELLSRSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIHLIAKKSKN  328

Query  61   HIQWL  65
            HIQWL
Sbjct  329  HIQWL  333
```

>ref|XP_874289.1| PREDICTED: similar to E2F transcription factor 2 [Bos taurus] Length=367 (**BtE2f3**)

Score = 105 bits (262), Expect = 5e-24
Identities = 51/65 (**78%**), Positives = 58/65 (**89%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY+TSL L TK+F+ LLS S DGV+DLNWAAEVL VQKRRIYDITNVLEGIQLI KK+KN
Sbjct  59   RYDTSLGLLTKKFIYLLSESEDGVLDLNWAAEVLDVQKRRIYDITNVLEGIQLIRKKAKN  118

Query  61   HIQWL  65
            +IQW+
Sbjct  119  NIQWV  123
```

>ref|XP_614932.2| PREDICTED: similar to E2F-3 transcription factor [Bos taurus] Length=463 (**BtE2f3**)

Score = 103 bits (256), Expect = 2e-23
Identities = 50/65 (**76%**), Positives = 58/65 (**89%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY+TSL L TK+F++LLS S DGV+DLN AAEVLKVQKRRIYDITNVLEGI LI KKSKN
Sbjct  176  RYDTSLGLLTKKFIQLLSQSPDGVLDLNKAAEVLKVQKRRIYDITNVLEGIHLIKKKSKN  235

Query  61   HIQWL  65
            ++QW+
Sbjct  236  NVQWM  240
```

>ref|XP_587085.3| PREDICTED: similar to transcription factor [Bos taurus] Length=314 (**BtTf1**)

Score = 80.9 bits (198), Expect = 1e-16
Identities = 43/65 (**66%**), Positives = 49/65 (**75%**), Gaps = 1/65 (1%)

```
Query  1   RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKV-QKRRIYDITNVLEGIQLIAKKSK  59
           R+E SL L T +F+ LL  + DGV+DL  AA+ L V QKRRIYDITNVLEGI LI KKSK
Sbjct  18  RHEKSLGLLTAKFVSLLQEAKDGVLDLKAAADTLAVRQKRRIYDITNVLEGIDLIEKKSK  77

Query  60  NHIQW  64
           N IQW
Sbjct  78  NSIQW  82
```

>ref|NP_001069341.1| hypothetical protein LOC525428 [Bos taurus] Length=404 (**BtP1**)

 Score = 80.5 bits (197), Expect = 2e-16
 Identities = 43/65 (**66%**), Positives = 49/65 (**75%**), Gaps = 1/65 (1%)

```
Query  1   RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKV-QKRRIYDITNVLEGIQLIAKKSK  59
           R+E SL L T +F+ LL  + DGV+DL  AA+ L V QKRRIYDITNVLEGI LI KKSK
Sbjct  17  RHEKSLGLLTTKFVSLLQEAKDGVLDLKLAADTLAVRQKRRIYDITNVLEGIGLIEKKSK  76

Query  60  NHIQW  64
           N IQW
Sbjct  77  NSIQW  81
```

>ref|XP_585927.2| PREDICTED: similar to E2F6 splice [Bos taurus] Length=288 (**BtE2f6S**)

 Score = 78.6 bits (192), Expect = 6e-16
 Identities = 37/65 (**56%**), Positives = 50/65 (**76%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            R  +SL+  T RF+ LL  S +GV+DLN AAE L + KRR+YD+TNVL GI+L+ KKS++
Sbjct  64   RCNSSLSDLTPRFMALLRSSPEGVLDLNKAAEALGIPKRRLYDVTNVLSGIKLVEKKSRS  123

Query  61   HIQWL  65
            HIQW+
Sbjct  124  HIQWI  128
```

>ref|NP_001070316.1| E2F transcription factor 6 [Bos taurus] (**BtE2f6**)
Length=285

 Score = 77.0 bits (188), Expect = 2e-15
 Identities = 34/65 (**52%**), Positives = 50/65 (**76%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            R++ SL   T++F++L+  +  G++DLN  A  L V+KRR+YDITNVL+GI L+ KKSKN
Sbjct  63   RFDVSLVYLTRKFMDLVRSAPGGILDLNKVATKLGVRKRRVYDITNVLDGIDLVEKKSKN  122

Query  61   HIQWL  65
            HI+W+
Sbjct  123  HIRWI  127
```

>ref|XP_001254261.1| PREDICTED: similar to E2F family member 8 [Bos taurus]
Length=1005 (**BtE2f8**)

 Score = 47.0 bits (110), Expect = 2e-06
 Identities = 26/68 (**38%**), Positives = 39/68 (**57%**), Gaps = 4/68 (5%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVD----LNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
            R E SL L   +FL    +  +  V+    L+  AE L V++RRIYDI NVLE + ++++
Sbjct  252  RKEKSLGLLCHKFLARYPNYPNPAVNNDICLDEVAEELNVERRRIYDIVNVLESLHMVSR  311

Query  57   KSKNHIQW  64
             +KN   W
Sbjct  312  LAKNRYTW  319
```

 Score = 35.8 bits (81), Expect = 0.005
 Identities = 23/70 (**32%**), Positives = 35/70 (**50%**), Gaps = 14/70 (20%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVL--------------KVQKRRIYDITN  46
```

```
            R + SL + +++F+ L   S   +V L  AA++L              K + RR+YDI N
Sbjct  400  RKDKSLKVMSQKFVTLFLVSTPQIVSLEIAAKILTWEDHVEDLDRSKFKTKIRRLYDIAN  459

Query  47   VLEGIQLIAK  56
            VL  + LI K
Sbjct  460  VLSSLDLIKK  469
```

>ref|XP_604488.3| PREDICTED: similar to E2F transcription factor 7 [Bos taurus]
Length=808 (**BtE2f7**)

```
 Score = 43.1 bits (100),  Expect = 3e-05
 Identities = 24/68 (35%), Positives = 37/68 (54%), Gaps = 4/68 (5%)

Query  1    RYETSLNLTTKRFLELLSH----SADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
            R + SL L  ++FL         +    + L+  A  L V++RRIYDI NVLE + L+++
Sbjct  102  RKQKSLGLLCQKFLARYPSYPLSTEKTTISLDEVAVSLGVERRRIYDIVNVLESLHLVSR  161

Query  57   KSKNHIQW  64
             +KN   W
Sbjct  162  VAKNQYSW  169


 Score = 35.8 bits (81),  Expect = 0.005
 Identities = 23/69 (33%), Positives = 36/69 (52%), Gaps = 13/69 (18%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVL-------------KVQKRRIYDITNV  47
            R + SL + +++F+ L   S   +V L+ AA++L              K + RR+YDI NV
Sbjct  242  RKDKSLKIMSQKFVMLFLVSKTKIVTLDVAAKILIEESQDIPDHSKFKTKVRRLYDIANV  301

Query  48   LEGIQLIAK  56
            L  + LI K
Sbjct  302  LTSLMLIKK  310
```

Rattus norvegicus

>ref|XP_001065036.1| PREDICTED: similar to Transcription factor E2F1 (E2F-1) [Rattus norvegicus] Length=528 (**RnE2f1a**)

 Score = 131 bits (329), Expect = 1e-31
 Identities = 65/65 (**100%**), Positives = 65/65 (**100%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN
Sbjct  221  RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  280

Query  61   HIQWL  65
            HIQWL
Sbjct  281  HIQWL  285
```

>ref|XP_230765.4| PREDICTED: similar to Transcription factor E2F1 (E2F-1) [Rattus norvegicus] Length=432 (**RnE2f1b**)

 Score = 131 bits (329), Expect = 1e-31
 Identities = 65/65 (**100%**), Positives = 65/65 (**100%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN
Sbjct  125  RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  184

Query  61   HIQWL  65
            HIQWL
Sbjct  185  HIQWL  189
```

>ref|XP_001078179.1| PREDICTED: similar to Transcription factor E2F3 (E2F-3) [Rattus norvegicus] (**RnE2f3**)
Length=417

 Score = 103 bits (256), Expect = 4e-23
 Identities = 50/65 (**76%**), Positives = 58/65 (**89%**), Gaps = 0/65 (0%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            RY+TSL L TK+F++LLS S DGV+DLN AAEVLKVQKRRIYDITNVLEGI LI KKSKN
Sbjct  353  RYDTSLGLLTKKFIQLLSQSPDGVLDLNKAAEVLKVQKRRIYDITNVLEGIHLIKKKSKN  412

Query  61   HIQWL  65
            ++QW+
Sbjct  413  NVQWM  417
```

>ref|XP_574892.1| PREDICTED: similar to Transcription factor E2F5 (E2F-5) [Rattus norvegicus] (**RnE2f5**)
Length=338

 Score = 80.9 bits (198), Expect = 2e-16
 Identities = 43/65 (**66%**), Positives = 49/65 (**75%**), Gaps = 1/65 (1%)

```
Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKV-QKRRIYDITNVLEGIQLIAKKSK  59
            R+E SL L T +F+ LL  + DGV+DL  AA+ L V QKRRIYDITNVLEGI LI KKSK
Sbjct  43   RHEKSLGLLTTKFVSLLQEAQDGVLDLKAAADTLAVRQKRRIYDITNVLEGIDLIEKKSK  102

Query  60   NHIQW  64
            N IQW
```

```
Sbjct  103  NSIQW  107
```

>ref|XP_233986.2| PREDICTED: similar to E2F transcription factor 6 [Rattus norvegicus] Length=290 (**RnE2f6**)

```
 Score = 77.8 bits (190),  Expect = 2e-15
 Identities = 34/65 (52%), Positives = 51/65 (78%), Gaps = 0/65 (0%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            R++ SL   T++F++L+  +  G++DLN  A  L V+KRR+YDITNVL+GI+L+ KKSKN
Sbjct  63   RFDVSLVYLTRKFMDLVRSAPGGILDLNKVATKLGVRKRRVYDITNVLDGIELVEKKSKN  122

Query  61   HIQWL  65
            HI+W+
Sbjct  123  HIRWI  127
```

>ref|XP_001069459.1| PREDICTED: similar to E2F transcription factor 6 isoform 1 [Rattus norvegicus] Length=237 (**RnE2f61**)

```
 Score = 77.8 bits (190),  Expect = 2e-15
 Identities = 34/65 (52%), Positives = 51/65 (78%), Gaps = 0/65 (0%)

Query  1   RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
           R++ SL   T++F++L+  +  G++DLN  A  L V+KRR+YDITNVL+GI+L+ KKSKN
Sbjct  28  RFDVSLVYLTRKFMDLVRSAPGGILDLNKVATKLGVRKRRVYDITNVLDGIELVEKKSKN  87

Query  61  HIQWL  65
           HI+W+
Sbjct  88  HIRWI  92
```

>ref|XP_001069501.1| PREDICTED: similar to E2F transcription factor 6 isoform 2 [Rattus norvegicus] Length=272 (**RnE2f62**)

```
 Score = 77.8 bits (190),  Expect = 2e-15
 Identities = 34/65 (52%), Positives = 51/65 (78%), Gaps = 0/65 (0%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAKKSKN  60
            R++ SL   T++F++L+  +  G++DLN  A  L V+KRR+YDITNVL+GI+L+ KKSKN
Sbjct  63   RFDVSLVYLTRKFMDLVRSAPGGILDLNKVATKLGVRKRRVYDITNVLDGIELVEKKSKN  122

Query  61   HIQWL  65
            HI+W+
Sbjct  123  HIRWI  127
```

>ref|XP_001053974.1| PREDICTED: similar to Transcription factor E2F5 (E2F-5) [Rattus norvegicus] Length=372 (**RnE2f5**)

```
 Score = 55.1 bits (131),  Expect = 1e-08
 Identities = 29/36 (80%), Positives = 30/36 (83%), Gaps = 1/36 (2%)

Query  30   AAEVLKV-QKRRIYDITNVLEGIQLIAKKSKNHIQW  64
            AA+ L V QKRRIYDITNVLEGI LI KKSKN IQW
Sbjct  106  AADTLAVRQKRRIYDITNVLEGIDLIEKKSKNSIQW  141
```

>ref|XP_001080267.1| PREDICTED: similar to E2f family member 8 isoform 2 [Rattus norvegicus] Length=877 (**RnE2f82**)

```
 Score = 46.2 bits (108),  Expect = 6e-06
 Identities = 26/68 (38%), Positives = 38/68 (55%), Gaps = 4/68 (5%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVD----LNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
            R E SL L   +FL       +  V+    L+  AE L V++RRIYDI NVLE + ++++
Sbjct  113  RKEKSLGLLCHKFLARYPKYPNPAVNNDICLDEVAEELNVERRRIYDIVNVLESLHMVSR  172

Query  57   KSKNHIQW  64
             +KN   W
Sbjct  173  LAKNRYTW  180


>ref|XP_001080259.1|   PREDICTED:  similar  to  E2f  family  member  8  isoform  1  [Rattus
norvegicus] (RnE2f81)
Length=877

 Score = 46.2 bits (108),  Expect = 6e-06
 Identities = 26/68 (38%), Positives = 38/68 (55%), Gaps = 4/68 (5%)

Query  1    RYETSLNLTTKRFLELLSHSADGVVD----LNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
            R E SL L   +FL       +  V+    L+  AE L V++RRIYDI NVLE + ++++
Sbjct  113  RKEKSLGLLCHKFLARYPKYPNPAVNNDICLDEVAEELNVERRRIYDIVNVLESLHMVSR  172

Query  57   KSKNHIQW  64
             +KN   W
Sbjct  173  LAKNRYTW  180


>ref|XP_001080823.1|    PREDICTED:  similar  to  E2F  transcription  factor  7  [Rattus
norvegicus] Length=1295 (RnE2f7)

 Score = 42.4 bits (98),  Expect = 8e-05
 Identities = 24/68 (35%), Positives = 37/68 (54%), Gaps = 4/68 (5%)

Query  1    RYETSLNLTTKRFLELLSH----SADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
            R + SL L  ++FL         +    + L+  A  L V++RRIYDI NVLE + L+++
Sbjct  536  RKQKSLGLLCQKFLARYPSYPLSTEKTTISLDEVAVSLGVERRRIYDIVNVLESLHLVSR  595

Query  57   KSKNHIQW  64
             +KN   W
Sbjct  596  VAKNQYGW  603


>ref|XP_235118.4|  PREDICTED: similar to E2F transcription factor 7 [Rattus norvegicus]
Length=1200 (RnE2f7b)

 Score = 42.4 bits (98),  Expect = 8e-05
 Identities = 24/68 (35%), Positives = 37/68 (54%), Gaps = 4/68 (5%)

Query  1    RYETSLNLTTKRFLELLSH----SADGVVDLNWAAEVLKVQKRRIYDITNVLEGIQLIAK  56
            R + SL L  ++FL         +    + L+  A  L V++RRIYDI NVLE + L+++
Sbjct  441  RKQKSLGLLCQKFLARYPSYPLSTEKTTISLDEVAVSLGVERRRIYDIVNVLESLHLVSR  500

Query  57   KSKNHIQW  64
             +KN   W
Sbjct  501  VAKNQYGW  508


>ref|XP_218601.4|  PREDICTED: similar to E2f family member 8 [Rattus norvegicus]
Length=992 (RnE2f8)

 Score = 38.1 bits (87),  Expect = 0.002
```

Identities = 14/29 (**48%**), Positives = 22/29 (**75%**), Gaps = 0/29 (0%)

```
Query  36   VQKRRIYDITNVLEGIQLIAKKSKNHIQW  64
            V++RRIYDI NVLE + ++++ +KN   W
Sbjct  267  VERRRIYDIVNVLESLHMVSRLAKNRYTW  295
```